A Profit-Based Measure of Lending Discrimination

Madison Coots
Harvard University
mcoots@g.harvard.edu

Robert Bartlett
Stanford University
rbartlett@law.stanford.edu

Julian Nyarko
Stanford University
jnyarko@law.stanford.edu

Sharad Goel
Harvard University
sgoel@hks.harvard.edu

Abstract

Algorithmic lending has transformed the consumer credit landscape, with complex machine learning models now commonly used to make or assist underwriting decisions. To comply with fair lending laws, these algorithms typically exclude legally protected characteristics, such as race and gender. Yet algorithmic underwriting can still inadvertently favor certain groups, prompting new questions about how to audit lending algorithms for potentially discriminatory behavior. Building on prior theoretical work, we introduce a profit-based measure of lending discrimination in loan pricing. Applying our approach to approximately 80,000 personal loans from a major U.S. fintech platform, we find that loans made to men and Black borrowers yielded lower profits than loans to other groups, indicating that men and Black applicants benefited from relatively favorable lending decisions. We trace these disparities to miscalibration in the platform's underwriting model, which underestimates credit risk for Black borrowers and overestimates risk for women. We show that one could correct this miscalibration—and the corresponding lending disparities—by explicitly including race and gender in underwriting models, illustrating a tension between competing notions of fairness.

1 Introduction

Fifty years ago, when the Fair Housing Act (FHA) (41) and the Equal Credit Opportunity Act (ECOA) (42) established the foundation for fair lending regulation, the lending landscape looked very different from today. At that time, borrowers typically applied for loans in person at a bank, where approval and pricing decisions were made at the discretion of

individual loan officers. Early fair lending regulation focused primarily on curbing discrimination arising from this individual discretion, particularly when influenced by implicit or explicit biases against racial minorities and women.

Over the last twenty years, however, the lending landscape has been transformed by the rise of algorithmic lending—and "fintech lending" in particular. Under the fintech model of lending, borrowers apply for loans online, and underwriting is handled by sophisticated machine learning models rather than individual lending officers. The rapid growth of fintech lending has been accelerated by promises of reduced subjectivity in lending decisions from eliminating human inputs, as well as cost-effective gains in accurately predicting creditworthiness by using highly complex underwriting models (17; 18). However, this shift from human to algorithmic decision making has increased concerns about *algorithmic* bias in the models used to make lending decisions (2; 6; 22; 28; 29).

These concerns about algorithmic bias have prompted scholars to revisit questions about how best to detect and regulate discrimination in the fintech era (5; 13; 25; 31; 32), building upon a large body of empirical research on detecting lending discrimination more generally (10; 11; 19; 24; 34; 35). In the most common regulatory approach, auditors consider "adverse impact ratios" (AIRs): the ratio of the decision rate (e.g., loan approval or interest rate) for members of a protected group relative to a reference group, such as men or non-Hispanic White borrowers. However, relatively low approval rates or high interest rates may simply reflect differences in creditworthiness between groups, rather than discrimination, limiting the probative value of adverse impact ratios.

Recognizing this limitation of AIRs, researchers have sought to estimate differences in approval rates and interest rates across groups after adjusting for observable risk factors in a regression model (7; 10; 11). This approach, however, likewise suffers from significant limitations, both practical and conceptual. In practice, researchers and regulators rarely have access to the full set of relevant information used by lenders to make approval or pricing decisions. Failing to adjust for this complete set of information can produce misleading results and may over- or understate discrimination (3). This potential for omitted-variable bias is even greater with algorithmic lending, where underwriting models often rely on thousands of proprietary covariates. Further, this regression-based approach can at best detect the *direct* effects of protected attributes on decisions, limiting its utility for auditing facially neutral lending models that do not explicitly condition decisions on race or gender (30). In this case, auditors that adjust for the full set of risk variables would (correctly) find no marginal effect of race or gender on decisions. Yet even with facially neutral algorithms, lenders may inadvertently discriminate against certain groups. For example, systematic er-

rors in estimating risk may lead lenders to favor or disfavor members of certain groups—an algorithmic variant of redlining (16).

Here we address these limitations by developing a simple and easily implementable outcome-based measure for discrimination in loan pricing, as defined under U.S. fair lending law. As a matter of U.S. federal law, fair-lending obligations prohibiting discriminatory practices arise principally under the Fair Housing Act (FHA) and the Equal Credit Opportunity Act (ECOA). U.S. courts have determined these statues prohibit two distinct types of discriminatory conduct (40). For one, they render illegal "disparate treatment," which encompasses lending decisions that are intentionally based on a protected characteristic, such as race or gender. Additionally, these laws go further and also prohibit certain facially neutral lending practices that create disparities in ways that correlate with protected characteristics. This latter form of discrimination is referred to as "disparate impact," and it is the focus of our analysis. To state a disparate-impact claim, a plaintiff must first identify a specific policy or practice that produces a statistically significant disparity between protected and unprotected groups. This showing shifts the burden to the defendant-lender to demonstrate that the challenged practice is necessary to achieve a legitimate and nondiscriminatory objective. Importantly, the only type of business objective that courts have recognized as sufficient to justify disparities in lending outcomes is creditworthiness.[1] Justifications stemming from "market forces"—such as differences in the existence of credit deserts or other circumstances affecting a borrower's willingness to pay—have explicitly been deemed illegitimate reasons incapable of justifying a difference in loan pricing.[2] Consistent with this legal framework, we define discriminatory loan pricing as pricing disparities that correlate with protected characteristics and that cannot be explained by differences in creditworthiness.

Our method for detecting discriminatory lending relies on straightforward intuition, first articulated by (8; 9): If a lender prices loans accurately, then—after setting interest rates to reflect risk—we would expect loans to be similarly profitable across groups. Indeed, in a competitive marketplace comprised of risk-neutral lenders, we would expect loans to be priced to achieve the same expected return for every individual borrower. Consequently,

[1]See A.B. & S. Auto Service, Inc. v. South Shore Bank of Chicago, 962 F.Supp. 1056 (N.D. Ill. 1997) ("[In a disparate impact claim under the ECOA], once the plaintiff has made the prima facie case, the defendant-lender must demonstrate that any policy, procedure, or practice has a manifest relationship to the creditworthiness of the applicant. . . "); Lewis v. ACB Business Services, Inc., 135 F.3d 389, 406 (6th Cir. 1998) ("The [ECOA] was only intended to prohibit credit determinations based on 'characteristics unrelated to creditworthiness.").

[2]See Miller v. Countrywide Bank, NA, 571 F.Supp.2d 251, 258 (D. Mass 2008) (rejecting argument that discrimination in loan terms among African American and White borrowers was justified as the result of competitive "market forces," noting that prior courts had rejected the "market forces" argument insofar that it would allow the pricing of consumer loans to be "based on subjective criteria beyond creditworthiness."

if loans made to a particular group are systematically more profitable, this suggests that the lender priced those borrowers too aggressively relative to their true risk, consistent with discrimination. To compare profitability, we rely on the annualized internal rate of return (IRR), a common measure of profit in lending that directly connects loan pricing with repayment outcomes. A key advantage of this approach is that it only requires access to repayment data—and, in particular, does not require knowledge of either the inputs or outputs of a lender's internal risk models. In practice, lenders are not perfectly risk neutral, and may expect a premium to lend to riskier borrowers. We return to this point below and extend our analysis to account for this possibility.

We apply our approach to data drawn from a large online financial technology platform, estimating the annualized IRR across racial and ethnic and gender groups. We find that loans made to Black borrowers are less profitable than those to other racial subgroups; and that loans made to men are less profitable than those made to women. These results suggest the lender's decision algorithm systematically benefits Black borrowers and men. We confirm this apparent benefit by tracing these disparities back to miscalibration in the lender's underwriting model, which underestimates the borrowing risk of Black borrowers and overestimates the risk of women. One can correct this miscalibration by explicitly including race and gender in risk models, but doing so would violate fair lending laws, illustrating a tension between competing notions of fairness.

2 Lending Data

We use data from a large online financial technology platform that facilitates lending to individuals for a variety of purposes. The platform employs a proprietary underwriting model that utilizes both traditional and non-traditional information to assess lending risk and originate loans. Both the loan approval and pricing decisions are fully automated, with no human intervention.

The platform provided two distinct, non-overlapping sets of data. The first dataset consists of 79,251 funded 3-year loans from 2019 (i.e., loans that were approved and originated). The second dataset consists of 125,345 3-year loan applications from 2021 to 2023 (i.e., a mix of both approved and denied loan applications, as well as loans approved by the lender, but not accepted by the applicant). For each individual in the data, we observe the application date, as well as the approval status and terms (amount and APR) of each loan for which the applicant was considered. We additionally observe the full set of proprietary risk variables used by the lender to estimate default risk at the time of application. For the set of funded

loans only, we observe the target returns (i.e., profit targets) set by the lender for each loan, and the borrower's monthly loan repayments.

Lenders are legally prohibited from collecting information on applicants' protected characteristics and so, in line with regulatory practices (14), applicant's demographics are inferred from their names. Specifically, the platform used the Bayesian Improved Surname Geocoding (BISG) proxy methodology (20; 21) to estimate the probability that each individual was White, Black, Hispanic, Asian, or "other." Gender estimates were similarly derived from applicants' first names. In our main analysis, we use the inferred BISG and gender probabilities as individual-level weights when taking population and group averages. In the Appendix, we describe robustness checks that reproduce our main results after instead labeling individuals with their single, most likely race and gender category, in line with recent recommendations in the fair lending literature (43). The results from this alternative analysis qualitatively mirror those from our primary approach.

3 Estimates of Profit by Race and Gender

To operationalize our test of discrimination, we compute the realized profit of the aggregate loan portfolio of each race and gender subgroup. Specifically, for each subgroup, we first combine the individual loan repayment histories (using the BISG race and gender probability estimates as weights) into an aggregate cash flow vector that indicates the net amount the lender received from borrowers in each group in each month. Vector entries may be either positive or negative—with negative entries indicating net loan dispersal that period.

To measure profits by race and gender, we compute the annualized internal rate of return (IRR) of each aggregate cash flow, a metric commonly used by lenders to characterize loan profitability. The annualized IRR is the annualized rate that makes the net present value of all loan repayments equal to the loan principal disbursed in the first period. IRR is scale-invariant (i.e., its magnitude does not depend on the size of the loan or portfolio of loans) and it incorporates the time value of money. IRR can range from –100% (total loss) through 0% (break-even) to arbitrarily large positive values (the faster or larger the repayments relative to the principal, the higher the IRR).

We apply our test to the set of 79,251 3-year loans that were approved and originated in 2019. In Figure 1, we plot the resulting estimates of profit by subgroup. We find that loans made to White, Hispanic, and Asian borrowers earn similar returns (between 8.3% and 8.8%), whereas loans made to Black borrowers earn noticeably less (7.7%). Likewise, the

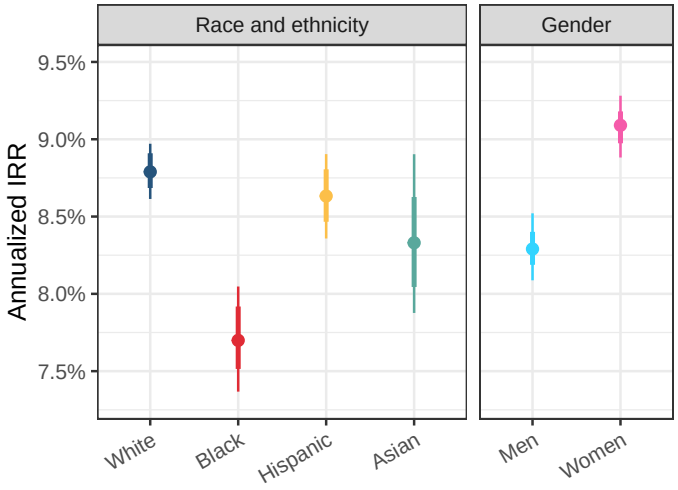

Figure 1: *Results of computing the annualized IRR of cashflows aggregated across race and gender groups. Relative to other groups, the lender earns less profit on loans made to Black borrowers and men, suggesting these groups benefit from relatively favorable loan terms. We show the 68% confidence interval (thick bar) and the 95% confidence interval (thin bar) of the IRR estimates.*

profits of loans made to men (8.3%) fall below those of women (9.1%). These results thus suggest the platform's algorithmic pipeline ultimately benefits Black borrowers and men, offering them relatively favorable loan terms.

We also operationalized our test a second way, by instead computing the IRR of individual loan cashflows and then taking a weighted average of individual IRRs using the race and gender probabilities as weights. Under this second approach, we estimate group profits by taking a weighted average of the IRR of individual loan cashflows within groups. Individual cash flows may have a mathematically undefined IRR if the cash flow reflects an immediate default (i.e., no payments are observed after the loan is disbursed) or an immediate prepayment (i.e., the loan is prepaid in the same period it is disbursed). In the case of immediate defaults, we set the IRR to be -100% to reflect a total loss. To address issues stemming from immediate prepayments, we constructed all loan cash flows so that the first observed payment occurs in the period immediately following loan disbursement. This way, a loan cannot be prepaid in the same period in which it was disbursed, and the resulting cash flow yields a well-defined IRR. The results of this approach for computing group profits are discussed further in Section 6, but are qualitatively similar to those shown in Figure 1.

Rather than examining differences in profit to identify discriminatory lending, one might instead consider differences in default rates. In particular, mirroring the logic of our profit-

based test, one might reason that a group of borrowers with relatively high observed default rates benefited from a less stringent lending standard, being granted loans even though they were relatively high risk. That approach, however, suffers from the problem of inframarginality (3; 23; 38). To see this, suppose that a lender perfectly estimates default risk for all applicants based on the available application data, and that loans are subsequently granted to all applicants with risk below a certain threshold, say 10%. In that scenario, lenders would be behaving efficiently—and would not be violating fair lending laws—yet, in general, we would expect group-specific default rates to differ. Indeed, the group-specific default rate is the conditional average of default risk for all members below the lending threshold. If risk distributions differ across groups, we would accordingly expect default rates among loan recipients to likewise differ, even in the absence of discrimination. Our profit-based measure, however, mitigates this concern. Assuming risk neutrality and a competitive marketplace, we would expect lenders to price loans to achieve identical expected returns from every applicant. If a lender's risk estimates are accurate, the average realized returns across subgroups would accordingly be similar, even if risk profiles differ across groups, sidestepping the problem of inframarginality (3).

4 Accounting for Risk Aversion

Our argument in favor of a profit-based test of discrimination rests critically on the assumption of risk neutrality. Potential risk aversion complicates the logic, since we expect a risk averse lender to demand a premium from higher risk borrowers. As a result, differences in risk across groups could lead to differences in IRR, even in the absence of discrimination—a manifestation of the problem of inframarginality.

Indeed, the lender we consider explicitly prices loans to achieve a target return that increases as a function of risk. Fig. 2 displays a sample target return curve provided to us by the lender, which operationalizes risk in terms of the "cumulative loss rate": the proportion of the principal that is expected to go uncollected from a prospective borrower. Formally, the cumulative loss rate is:

$$\frac{1}{P_0} \sum_t f(t) \cdot (P_t - R_t),$$

where P_0 is the original principal; for each period t, $f(t)$ is the probability of default in that period; P_t is the remaining principal; and R_t is the expected recovery conditional on default. The cumulative loss rate thus captures the likelihood and timing of defaults, as well as expected recovery amounts in the case of such events.

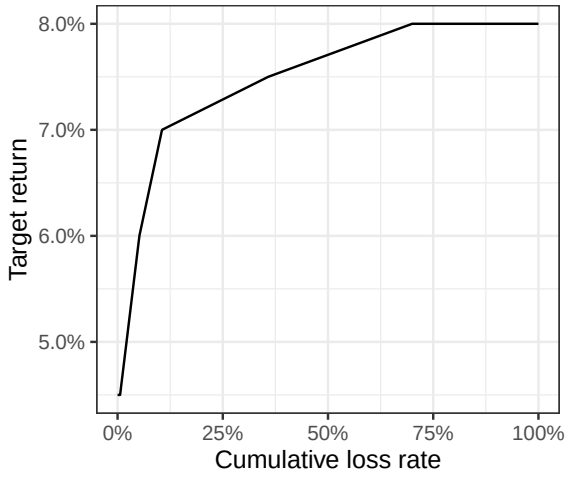

Figure 2: *An example target return curve used by the lender to price loans. The cumulative loss rate is the proportion of the principle that the lender expects to lose. The curve indicates a preference for relatively higher returns for riskier loans, consistent with risk aversion.*

In our analysis above, we observed that Black and male borrowers were less profitable. This pattern would be consistent with non-discrimination if it were also the case that these groups were lower risk than their non-Black and female counterparts—and thus the lender intentionally sought lower returns from them. We empirically investigate this possibility in two ways. First, we consider the lender's stated target return for each borrower. We display the average target returns across groups in the left panel of Figure 3, and find that Black and male borrowers are among the groups for which average target returns are highest—not lowest. Second, in the right panel of Figure 3, we compute the observed proportion of principal lost by group, similarly finding that Black and male borrowers are among the highest risk groups. (We assume for simplicity that $R_t = 0$, meaning that no additional principal was recovered after default.) It therefore appears that risk aversion does not explain the profit gaps we observe. If anything, we would expect, in the absence of discrimination, that Black and male borrowers would be slightly more, not less, profitable.[3]

[3]In practice, target returns may not be readily available to regulators to conduct the type of analysis we carryout above; and, even if they were available, regulators may not wish to rely on such lender attestations. While cumulative loss can be approximated by loan repayment data, regulators might instead consider using observed default rates as a simple, more easily interpreted measure of risk. Figure A5 in the Appendix shows that Black borrowers and men have higher default rates than other groups, in line with the cumulative loss comparisons. Moreover, Figure A6 shows that default rates and target returns are empirically strongly correlated, further illustrating the value of considering default rates when accounting for risk aversion. Regulators might also examine differences in the volatility of IRR across groups as an additional indicator of group-level risk, as shown in Figure A7 in the Appendix.

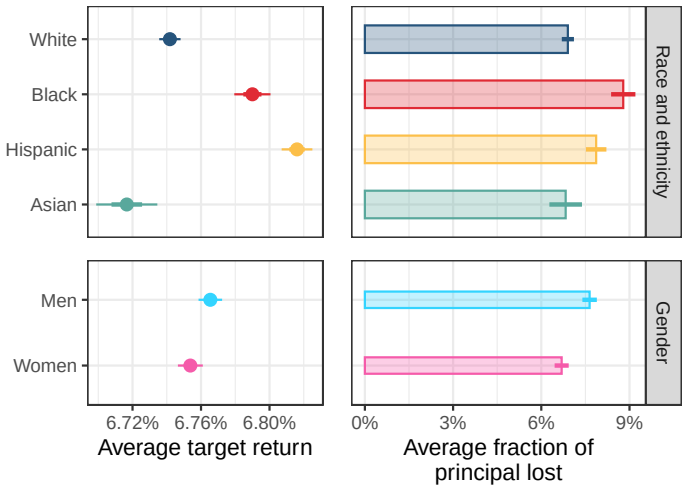

Figure 3: *Average target returns and the average fraction of principal lost across race and gender groups. (Left) The lender has higher target returns for Black, Hispanic, and men borrowers than other groups. Target returns are set by the lender as a function of the estimated cumulative loss rate for each loan, as shown in Figure 2. Differences in group average target returns therefore correspond to differences in the average estimated cumulative loss rates across groups. We show the 68% confidence interval (thick bar) and the 95% confidence interval (thin bar) of the estimates of average target returns within each group. (Right) The lender loses greater fractions of loan principals to Black borrowers and men. The thick line shows the 95% confidence interval of our estimates.*

5 Miscalibrated Risk Estimates

What then drives the differences in profitability across borrowers? One possibility is that the lender systematically underestimates the risk of Black and male borrowers, leading to more favorable interest rates relative to their true riskiness. To investigate this possibility, we would ideally assess the calibration of the lender's internal risk score, i.e., comparing predicted and realized default rates. However, this risk score is proprietary and was withheld by the lender. Instead, we evaluate the calibration of our own (race- and gender-blind) risk model trained with the same set of proprietary features used by the lender's internal model.

Using the approximately 80,000 funded loans from 2019, we trained two risk models using XGBoost with 5-fold cross-validation to obtain out-of-sample predictions. Our first model—the "blind" model—serves as a proxy for the lender's internal underwriting model because it is race- and gender-blind, meaning that it does not use information on race and ethnicity or gender to make risk predictions. Our second model, however—the "aware" model—is race- and gender-aware and uses the BISG probabilities to estimate risk.

A calibrated underwriting model should have estimated default rates that align with observed default rates. As shown in Figure 4, our blind model is well-calibrated for White, Asian, and Hispanic borrowers, but is noticeably miscalibrated for Black borrowers, significantly *underestimating* Black borrowers' risk of default. Disaggregated by gender, we find that the model is in fact well-calibrated for men, but slightly miscalibrated for women, *overestimating* women's default risk.

By approximately reconstructing the lender's risk model, our findings suggest that miscalibration at least in part explains the observed differences in profitability across groups. But our analysis is inherently limited since we do not have access to the actual, proprietary risk model the lender employs. To address this limitation, we note that we would expect APR to increase approximately monotonically in the lender's internally estimated risk score. Leveraging this observation, we examine group-specific default rates as a function of APR in Figure 5. In the absence of any miscalibration, we would expect to see similar default rates across groups given a fixed APR. However, we instead find that, for any given APR, Black borrowers default at higher rates than White borrowers, and women borrowers default at lower rates than men. This result suggests that the miscalibration we observe in our own blind risk model likewise occurs in the lender's own internal model. This general strategy of assessing model miscalibration is one that regulators could also follow, since loan-level APR and default are typically readily available.

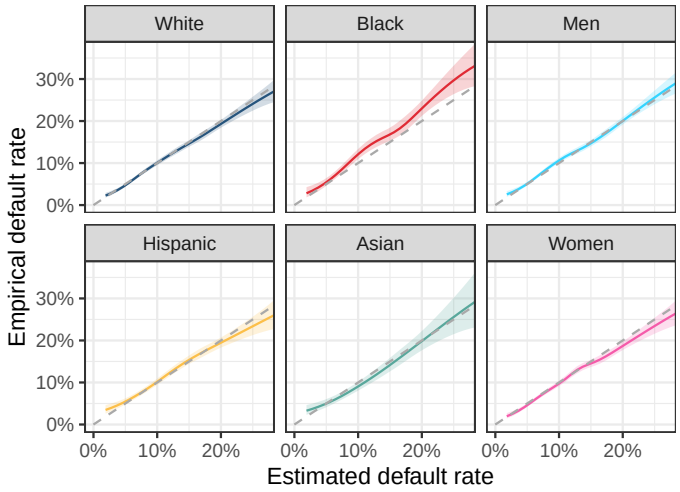

Figure 4: *The calibration of reconstructed race- and gender-blind risk scores, where the line $y = x$ denotes the line of perfect calibration, shown by a dashed gray line. (Left and Middle) When disaggregated across racial and ethnic groups, blind risk scores are miscalibrated for Black borrowers and tend to underestimate their likelihood of default. (Right) Blind risk scores are calibrated for men, but tend to slightly overestimate the risk of women. These results collectively suggest that inaccuracies in the lender's (blind) risk score lead to relative underpricing of loans for Black borrowers, and relative overpricing of loans for women.*

In general, miscalibration is not unexpected in race- and gender-blind models, since it is possible that groups exhibit differences in risk that remain even after adjusting for the factors included in the model (16). And our finding that the lender's risk model is miscalibrated for Black and women borrowers is consistent with past work that has observed similar patterns with other blind estimates of borrower risk, such as credit scores (4) and criminal risk assessments (39). One way to statistically correct this miscalibration is to explicitly include race and gender in risk models. In Figure 6, we plot the calibration of our race- and gender-aware model, and confirm that risk estimates align with observed rates of default for all groups.

We would consequently expect use of an aware risk model to eliminate the observed gaps in group profits, through changes in both loan approvals and pricing. To estimate changes in approval rates, we first trained a model to predict whether an applicant's first loan application was approved by the lender using the applicant's blind risk score. Specifically, using data on applicants who applied for 3-year loans, we fit a logistic regression model of the following form:

$$\Pr(\text{Approval}_i = 1) = \text{logit}^{-1}(\alpha + \beta R_{B,i}),$$

where Approval_i is a binary variable that indicates whether the first application for applicant

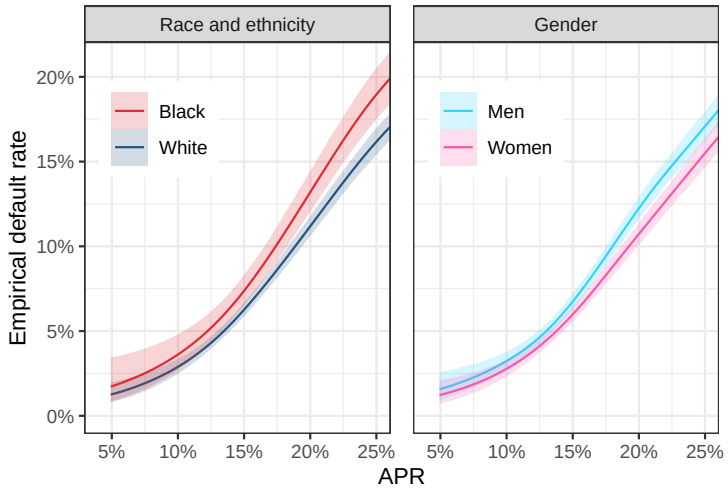

Figure 5: *An empirical assessment of the calibration of the lender's internal risk score. (Left) Given a fixed APR, Black borrowers default at higher rates than White borrowers. (Right) Similarly, given a fixed APR, women default at slightly lower rates than men. For each curve, we show the 95% confidence interval. Curves were smoothed using a generalized additive model using a logit link function. These results suggest that the lender's internal risk score exhibits miscalibration similar to that which we observe in our own blind risk model, resulting in relatively favorable APRs for Black borrowers relative to White borrowers, and slightly unfavorable APRs for women relative to men.*

i was approved, and $R_{B,i}$ is the blind risk score of applicant i with coefficient β. We then used our fitted approval model to obtain predictions of an applicant's probability of approval using their blind risk score $R_{B,i}$, as well as their aware risk score $R_{A,i}$. Finally, using the BISG predictions, we computed a weighted average of the difference between aware and blind predicted approval probabilities to obtain expected changes in approval rates within each group. In Figure 7, we show that under this aware model, approval rates would decrease for Black borrowers (-3pp) and men (-0.5pp), and correspondingly increase for other racial and ethnic groups ($\approx$ +1pp) and women (+1.5pp).

To estimate changes in APR, we took a similar approach and trained a model to predict the approved APR of the applicant's first loan application as a function of their blind risk score and the federal interest rate at the time of the application. Specifically, using data on applicants approved for 3-year loans, we fit a log-linear regression model of the following form:

$$\log(\text{APR}_i) = \alpha + \beta R_{B,i} + \gamma F_i,$$

where APR_i is the APR approved for applicant i, $R_{B,i}$ is the blind risk score of applicant i with coefficient β, and F_i is the federal interest rate at the time of applicant i's application with coefficient γ. We then used our APR model to obtain predictions of an applicants approved APR using their blind risk score $R_{B,i}$, as well as their aware risk score $R_{A,i}$. Using the BISG predictions, we computed a weighted average of the difference between aware and blind predicted APRs to obtain expected changes in APR within each group. As shown in the right panel of Figure 7, average APRs would remain relatively unchanged for White, Hispanic, Asian, and women borrowers, but would increase by 0.8 points for Black borrowers and 0.3 points for men. We include these results for illustrative purposes only, and note that the use of a race- or gender-aware underwriting model would be legally impermissible under current fair lending law.

6 Strategic Shopping

Finally, in addition to risk aversion and model miscalibration, we consider a third possible explanation for the gaps in profitability we see: strategic shopping. In particular, it may be the case that men and Black borrowers are more sophisticated shoppers and only accept loan offers that are relatively favorable. To test this theory, we make use of data on the initial offers presented to applicants. We then estimate counterfactual group IRR's in the absence of any strategic shopping—that is, assuming applicants accept any loan they were approved for. That is, applicants do not walk away from loans they have been approved for

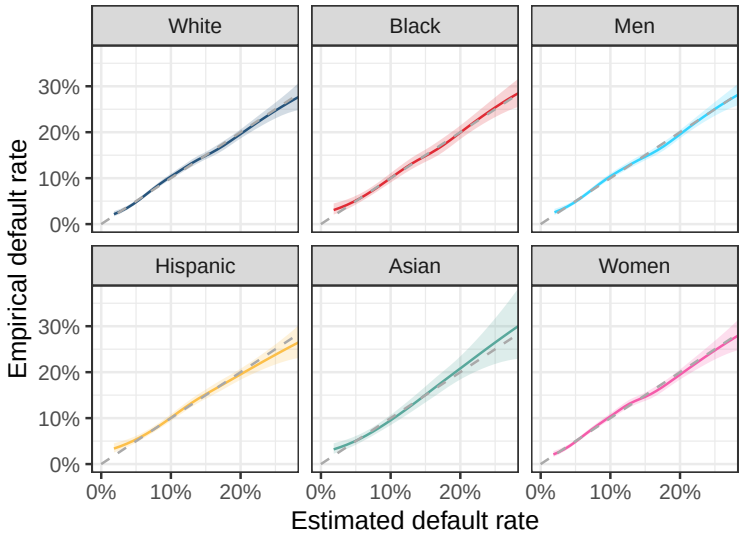

Figure 6: *The calibration of reconstructed race- and gender-blind risk scores, where the line $y = x$ denotes the line of perfect calibration, shown by a dashed gray line. (Left and Middle) Aware risk scores are calibrated for all racial and ethnic subgroups. (Right) Similarly, aware risk scores are calibrated for both men and women. For each curve, we show the 95% confidence interval. Curves were smoothed using a generalized additive model using a logit link function. These results suggest that an aware risk score would correct the inaccuracies in loan pricing stemming from the use of a blind risk score.*

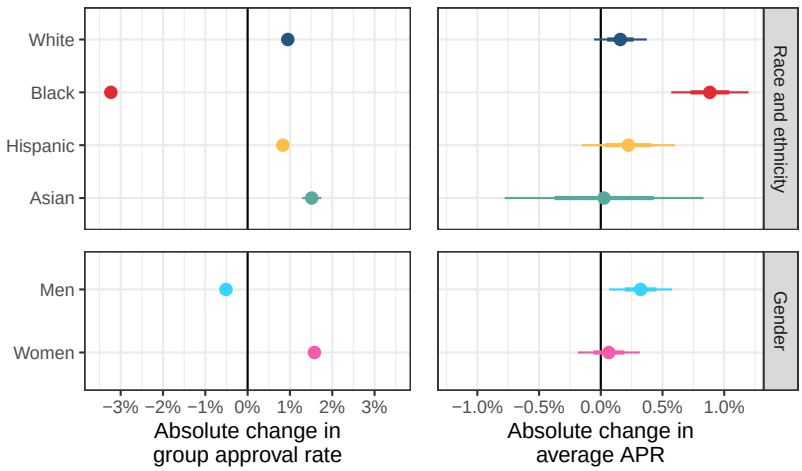

Figure 7: *The predicted consequences of switching to a race- and gender-aware risk model for loan underwriting. (Left) When disaggregated by race and ethnicity, loan approval rates would be expected to decrease for Black borrowers by more than 3pp, and increase by approximately 1pp for White, Hispanic, and Asian borrowers. By gender, approval rates would be expected to decrease for men by about 0.5pp and increase for women by approximately 1.5pp. (Right) Across racial and ethnic groups, loan APRs would increase for Black borrowers, but remain relatively unchanged for White, Hispanic, and Asian borrowers. By gender, APRs would increase for men, but remain constant for women. We show the 68% confidence interval (thick bar) and the 95% confidence interval (thin bar) for each estimate.*

or request and accept a counteroffer for a different loan amount or APR.

To determine the counterfactual group IRRs of this no-shopping scenario, we trained a model to predict the IRR of individual loans using the applicant's race-and gender-aware risk score, the loan amount, and APR. We used aware risk scores to mitigate model miscalibration, as our goal was to estimate outcomes, not inform lending decisions. We trained this model using data of funded borrowers for whom the loan amount they requested equaled the loan amount that they ultimately received. Specifically, we fit a linear regression model of the following form:

$$\text{IRR}_i = \alpha R_{A,i} + \gamma \text{LoanAmount}_i + \beta \text{APR}_i$$

where IRR_i is the realized annualized IRR of the loan made to applicant i, $R_{A,i}$ is the aware risk score for applicant i with coefficient α, LoanAmount_i is the amount of the loan made to applicant i with coefficient γ, and APR_i is the APR of the loan made to applicant i with coefficient β. Then, using data on applicants approved for 3-year loans, we obtained counterfactual IRR's by applying our fitted model, assuming that each approved applicant accepted the loan they initially requested. Finally, we then used the BISG predictions to compute a weighted average of individual loan IRRs within each group.

Figure 8 compares these estimated counterfactual IRRs (triangular points)—absent any shopping—to the real IRRs we observe for actually funded borrowers (circular points). Note that the real IRRs shown in Figure 8 differ slightly from those in Figure 1 because Figure 1 reports each group's aggregate portfolio IRR, whereas Figure 8 reports a weighted average of individual loan IRRs within each group (the second operationalization of our test). We used this second approach to facilitate a more direct comparison between the counterfactual and real group IRRs.

Figure 8 shows that the counterfactual IRRs mirror the real IRRs, with Black borrowers generating significantly lower profits than other racial and ethnic groups, and men generating lower profits than women. We note that the counterfactual IRRs are estimated to be lower than the real IRRs because, if every applicant were to accept their loan offer, the lender would get a lower risk portfolio without losing the best borrowers to other lenders. In reality, lower risk borrowers often walk away or shop for better terms, leaving the lender with a riskier set of borrowers and higher overall profits associated with the added risk they are taking on. Ultimately, the similar patterns between the counterfactual and real group IRRs suggest that the profit gaps we observe are not explained by group differences in shopping behavior.

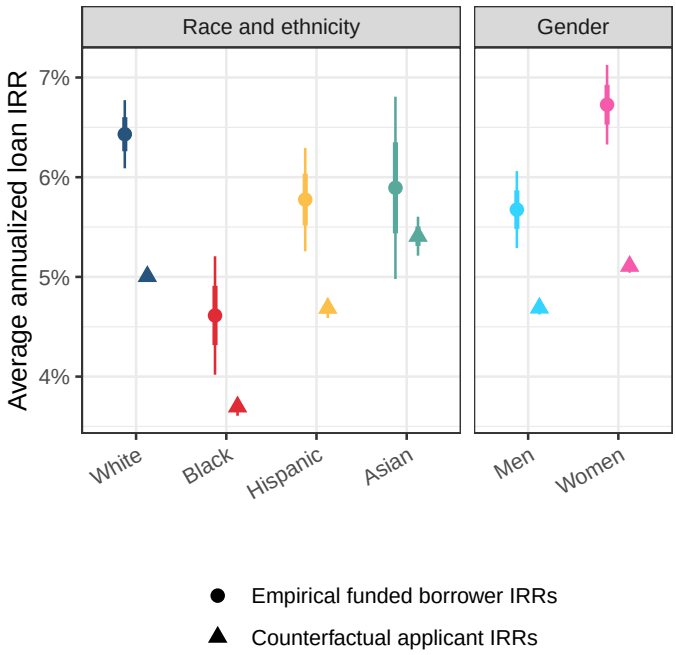

Figure 8: *Counterfactual and realized internal rates of return (IRRs). Counterfactual IRRs are estimated for loan applicants under the assumption that all approved applicants accept their first offer, removing effects of strategic borrower shopping and loan selection. Realized IRRs are calculated from the actual set of funded loans. Both sets of estimates are calculated by averaging IRRs of individual loans. Similar to the realized IRRs, the counterfactual IRRs of loans to men and Black borrowers are lower relative to other groups, suggesting that the profit gaps we observe are not solely attributable to potential differences in shopping behavior across groups.*

7 Discussion

In our analysis of data on approximately 80,000 loans originated by a major U.S. fintech platform, we found that loans to Black borrowers and men yielded lower profits than loans made to White, Hispanic, Asian and women borrowers. Our evidence suggests that this result is likely driven by miscalibration in the risk model used by the platform, which underestimates the default risk of Black borrowers and overestimates the default risk of women, leading the platform to inaccurately price loans for these groups. In turn, conditional on true riskiness, Black borrowers in our sample receive more favorable interest rates relative to other racial and ethnic groups, and women in our sample receive less favorable interest rates relative to men, on average.

Under fair lending statutes like the Equal Credit Opportunity Act of 1974, such risk-independent disparities could constitute a form of impermissible disparate impact. As we show, using a race- and gender-aware underwriting model could remove the gap, although doing so would constitute a form of legally impermissible race- and gender-based decision making, or disparate treatment (36). This tension echoes recent legal scholarship calling for regulatory frameworks better tailored to the oversight of algorithmic pricing systems used in credit markets (25; 26).

Our IRR-based test of discrimination is a practical and straightforward regulatory approach for monitoring lenders that employ algorithmic underwriting. This test aligns with recent work in the algorithmic fairness literature advocating for a shift towards outcome-based approaches to evaluating the fairness of algorithms (13; 25). It also addresses the well-known limitations of adverse impact ratios, the problem of omitted-variable bias in benchmark tests of discrimination (3), and the problem of inframarginality in tests based on default rates (38). Its simplicity—requiring only applicant demographics and loan cash flow vectors—also means that it can be easily generalized to other credit markets and serve as a useful first indicator of whether additional regulatory scrutiny is warranted.

While we believe our profit-based approach is broadly useful, it is subject to several important limitations. First, our test cannot detect all forms of potential discrimination. For example, a lender might hypothetically refuse to lend to a random subset of minority applicants, but then offer non-discriminatory loan terms to those who are ultimately offered loans. Such discriminatory behavior would not be detected by a comparison of profits across groups. Second, current legislation prohibits lenders from collecting protected characteristics of non-mortgage applicants, and so similar analyses would need to use imputed demographics, as we do here (42). It is worth noting, however, that there is precedent for

using BISG estimates by regulators, such as the Consumer Financial Protection Bureau and the Department of Justice (1; 14; 15). Finally, our analysis focused on comparing profits across individual racial and ethnic and gender groups, but recent work has suggested that the intersection of race and gender is an important dimension to consider when evaluating lending disparities across groups (27).

Our analysis focused on personal loans originated by a single U.S. fintech lender, but the findings likely generalize to other lenders and credit markets. For example, recent work uncovered similar miscalibration in consumer credit scores for Black borrowers (4) and other work observed underestimation of delinquency risk in business credit scores for minority business owners (37). Other recent work scrutinizing the calibration of consumer credit scores across gender in the subprime borrowing context similarly found that the default risk for women is overestimated (33). This pattern suggests that, in credit markets where these credit scores are used, other profit-based outcome tests could reveal analogous disparities to those we uncover in our analysis. Nevertheless, the magnitude and direction of profit disparities may vary across borrower populations and credit products, and replication of our analyses in different credit contexts would be a valuable next step in assessing the prevalence of the patterns we observe.

Our work illustrates that even facially neutral algorithms can create consequential racial and gender disparities, and underscores the importance of predictive accuracy in the pursuit of equitable lending (12). By developing and applying a profit-based test for discrimination, we hope to assist researchers and policymakers in the study, regulation, and remediation of lending discrimination.

Acknowledgments

M. Coots was supported by a James M. and Cathleen D. Stone PhD Scholar fellowship from the Stone Program in Wealth Distribution, Inequality, and Social Policy at Harvard University, and by a Social Equity and Health Equity stipend from the Malcolm Wiener Center for Social Policy at Harvard Kennedy School. The authors thank Desmond Ang, Susan Athey, Matthew Baum, Deirdre Bloome, Lucas Chancel, Adam Chilton, Johann Gaebler, Talia Gillis, Jacob Jameson, Taeku Lee, Daniel Schneider, Maya Sen, Max Spohn, and Richard Zeckhauser for helpful conversations and feedback.

References

[1] M. Akinwumi, J. Merrill, L. Rice, K. Saleh, and M. Yap. An ai fair policy lending agenda for the federal financial regulators, 2021. Brookings Center on Regulation and Markets.

[2] Robert B Avery, Kenneth P Brevoort, and Glenn Canner. Does credit scoring produce a disparate impact? *Real Estate Economics*, 40:S65–S114, 2012.

[3] Ian Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

[4] Trevor J Bakker, Stefanie DeLuca, Eric A English, James S Fogel, Nathaniel Hendren, and Daniel Herbst. Credit access in the united states. Technical report, National Bureau of Economic Research, 2025.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning*. 2019. URL `http://fairmlbook.org`.

[6] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104: 671, 2016.

[7] R. Bartlett, A. Morse, R. Stanton, and N. Wallace. Consumer-lending discrimination in the fintech era. *J Financ Econ*, 143(1):30–56, 2022. doi: 10.1016/j.jfineco.2021.05.031.

[8] G.S. Becker. *The economics of discrimination*. University of Chicago Press, 2nd edition, 1971. Originally published in 1957.

[9] G.S. Becker. The evidence against banks doesn't prove bias, 1993. In Bloomberg Businessweek (Online) (No. 3315, p. 18), Bloomberg Finance LP.

[10] N. Bhutta and A. Hizmo. Do minorities pay more for mortgages? *Rev Financ Stud*, 34(2):763–789, 2020.

[11] N. Bhutta, A. Hizmo, and D. Ringo. How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions. *The Journal of Finance*, LXXX(WP 24-09), 2024.

[12] Laura Blattner and Scott Nelson. How costly is noise? data and disparities in consumer credit. *arXiv preprint arXiv:2105.07554*, 2021.

[13] A. Chohlas-Wood, M. Coots, S. Goel, and J. Nyarko. Designing equitable algorithms. *Nature Computational Science*, 3(7):601–610, 2023.

[14] Consumer Financial Protection Bureau. Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment. Technical report, Consumer Financial Protection Bureau, Washington, D.C., 2014.

[15] Consumer Financial Protection Bureau and United States Department of Justice. Consent Order In the Matter of Ally Financial Inc. and Ally Bank, 2013. Administrative Proceeding File No. 2013-CFPB-0010.

[16] S. Corbett-Davies, J.D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. *J Mach Learn Res*, 24(312):1–117, 2023. URL `http://jmlr.org/papers/v24/22-1511.html`.

[17] M. Di Maggio and V. Yao. Fintech borrowers: Lax screening or cream-skimming? *Review of Financial Studies*, 34(10):4565–4618, 2020. doi: 10.1093/rfs/hhaa142.

[18] M. Di Maggio, D. Ratnadiwakara, and D. Carmichael. Invisible primes: Fintech lending with alternative data. Working Paper 29840, National Bureau of Economic Research, 2022. URL `http://www.nber.org/papers/w29840`.

[19] W. Dobbie, A. Liberman, D. Paravisini, and V. Pathania. Measuring bias in consumer lending. *Rev Econ Stud*, 88(6):2799–2832, 2021. doi: 10.1093/restud/rdaa078.

[20] M.N. Elliott, A. Fremont, P.A. Morrison, P. Pantoja, and N. Lurie. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*, 43:1772–1736, 2008.

[21] M.N. Elliott, P.A. Morrison, A. Fremont, D.F. McCaffrey, P. Pantoja, and N. Lurie. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Methodol*, 9:69–83, 2009.

[22] A Fuster, P Goldsmith-Pinkham, T Ramadorai, and A Walther. Predictably unequal? The effects of machine learning on credit markets, 2017.

[23] Johann D. Gaebler and Sharad Goel. A simple, statistically robust test of discrimination. *Proceedings of the National Academy of Sciences*, 122(10):e2416348122, 2025.

[24] M. Giacoletti, R. Heimer, and E.G. Yu. Using high-frequency evaluations to estimate discrimination: Evidence from mortgage loan officers, 2023. Available at SSRN 3795547.

[25] T. B. Gillis. The input fallacy. *Minnesota Law Review*, 106(3):1175, 2022. ISSN 0026-5535.

[26] Talia B. Gillis. "price discrimination" discrimination. Working paper, Columbia Law School; draft (Nov 2024), available on SSRN, 2024.

[27] Sarah K Harkness. Discrimination in lending markets: Status and the intersections of gender and race. *Social Psychology Quarterly*, 79(1):81–93, 2016.

[28] Mikella Hurley and Julius Adebayo. Credit scoring in the era of big data. *Yale JL & Tech.*, 18:148, 2016.

[29] Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin. The fairness of credit scoring models. *Management Science*, 2024.

[30] J. Jung, S. Corbett-Davies, J. Gaebler, R. Shroff, and S. Goel. Measuring disparate impact in human and machine decisions. *Proceedings of the National Academy of Science*, 2026. (forthcoming).

[31] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

[32] I. E. Kumar, K. E. Hines, and J. P. Dickerson. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with u.s. fair lending regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022. doi: 10.48550/arXiv.2210.02516.

[33] Zilong Liu and Hongyan Liang. Are credit scores gender-neutral? evidence of mis-calibration from alternative and traditional borrowing data. *Journal of Behavioral and Experimental Finance*, 47:101081, 2025. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2025.101081. URL `https://www.sciencedirect.com/science/article/pii/S2214635025000620`.

[34] A.H. Munnell, G.M.B. Tootell, L.E. Browne, and J. McEneaney. Mortgage lending in boston: Interpreting hmda data. *Am Econ Rev*, 86:25–53, 1996.

[35] K.A. Park. Measuring risk and access to mortgage credit with new disclosure data. *J Struct Finance*, 26(4):53–72, 2021.

[36] Devin G Pope and Justin R Sydnor. Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3(3):206–231, 2011.

[37] Alicia Robb and David T Robinson. Testing for racial bias in business credit scores. *Small Business Economics*, 50(3):429–443, 2018.

[38] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of inframarginality in outcome tests for discrimination. *Annals of Applied Statistics*, 11:1193–1216, 2017.

[39] Jennifer Skeem, John Monahan, and Christopher Lowenkamp. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40 (5):580, 2016.

[40] Winnie. F. Taylor. The ecoa and disparate impact theory: A historical perspective. *Journal of Law and Policty*, 26(2):575–635, 2018.

[41] U.S. Congress. Fair Housing Act, 1968.

[42] U.S. Congress. Equal Credit Opportunity Act, 1974.

[43] Yan Zhang. Assessing fair lending risks using race/ethnicity proxies. *Management Science*, 64(1):178–197, 2018.

Appendix

BISG Robustness Check

In line with recommendations from the Consumer Financial Protection Bureau, our main results were obtained by using the BISG probability estimates as weights throughout our analyses (14). However, more recent work has recommended using inferred labels on race and gender status by taking the argmax of BISG probability estimates (43). As a robustness check on our results, we regenerated our main results using this approach. Concretely, $R \in \{\text{Asian}, \text{Black}, \text{Hispanic}, \text{White}\}$ was assigned for each individual as

$$R = \arg\max_{R} \hat{p}_R$$

and $G \in \{\text{Woman}, \text{Man}\}$ was assigned for each individual as

$$G = \arg\max_{G} \hat{p}_G$$

where $\hat{p}_R$ and $\hat{p}_G$ denote the estimated probability distributions of membership in race and gender groups, respectively.

In Figures A1, A2, A3, and A4 we offer versions of Figures 1, 3, 4, and 8 generated using the above approach.

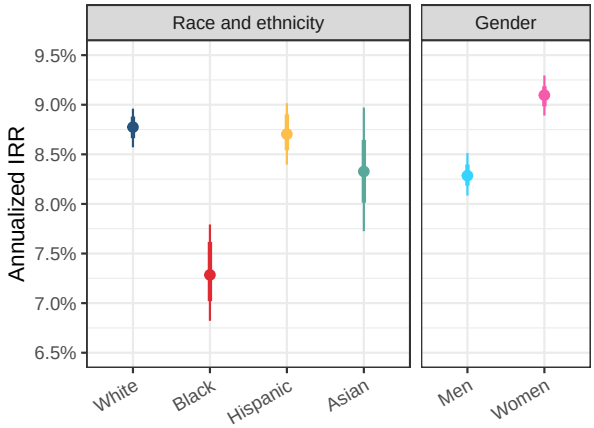

Figure A1: *Robustness check of Figure 1 using the argmax of BISG estimates to obtain race and gender labels, as opposed to using the BISG estimates as weights. The group IRR estimates produced under the argmax method are qualitatively similar to those shown in Figure 1 in the main text.*

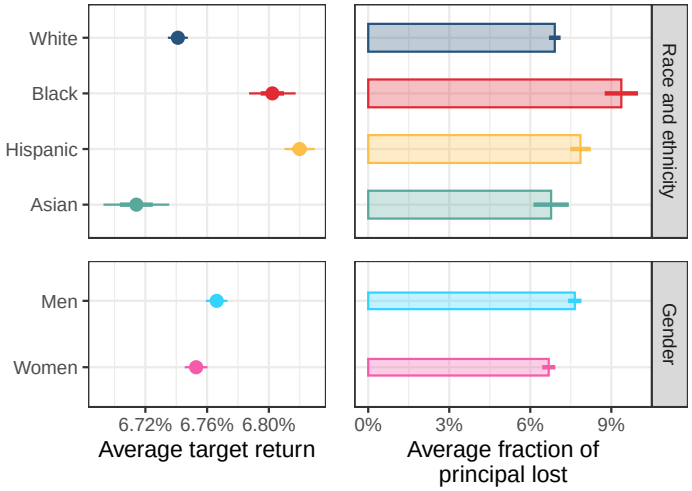

Figure A2: *Robustness check of Figure 3 using the argmax of BISG estimates to obtain race and gender labels, as opposed to using the BISG estimates as weights. Both sets of results produced under the argmax method are qualitatively similar to those shown in Figure 3 in the main text.*

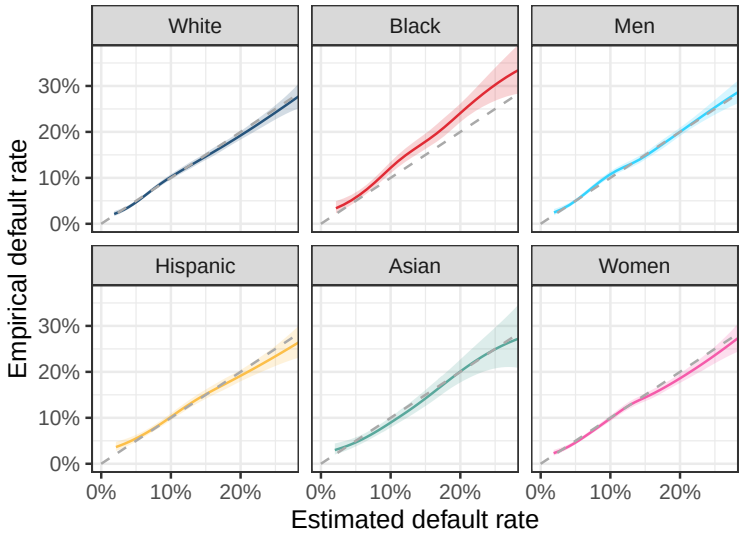

Figure A3: *Robustness check of Figure 4 using the argmax of BISG estimates to obtain race and gender labels, as opposed to using the BISG estimates as weights. The group calibration results produced under the argmax method are qualitatively similar to those shown in Figure 4 in the main text.*

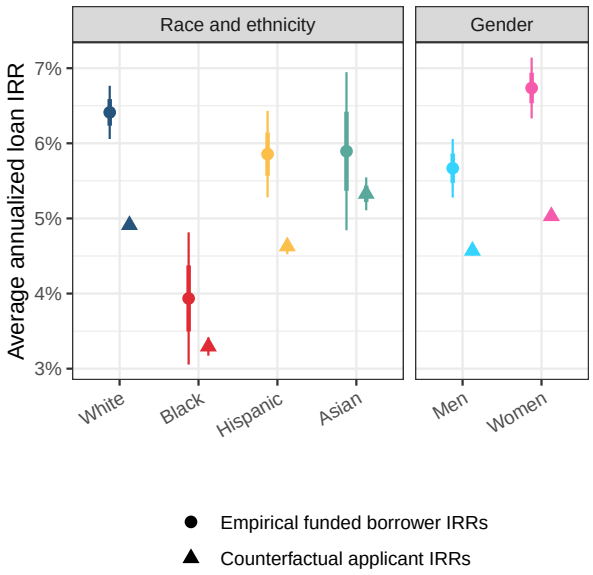

Figure A4: *Robustness check of Figure 8 using the argmax of BISG estimates to obtain race and gender labels, as opposed to using the BISG estimates as weights. These results produced under the argmax method are qualitatively similar to those shown in Figure 8 in the main text.*

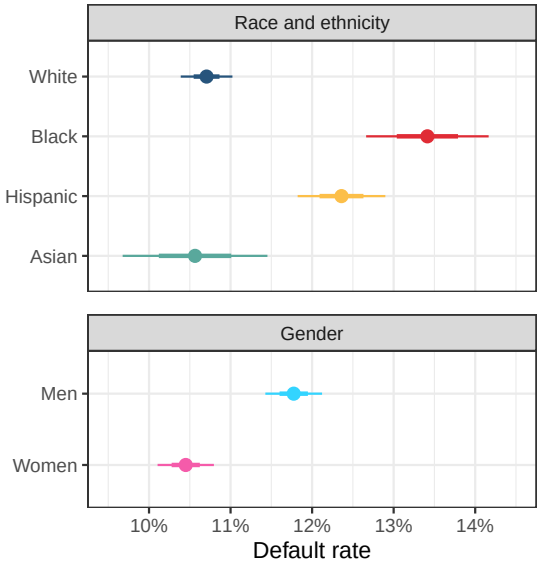

Figure A5: *Rates of loan default computed across race and gender groups. Black and male borrowers tend to be higher risk groups, as exhibited by their relatively higher rates of default. We show the 68% confidence interval (thick bar) and the 95% confidence interval (thin bar) of the default rate estimates. These estimates were computed using the set of funded borrowers.*

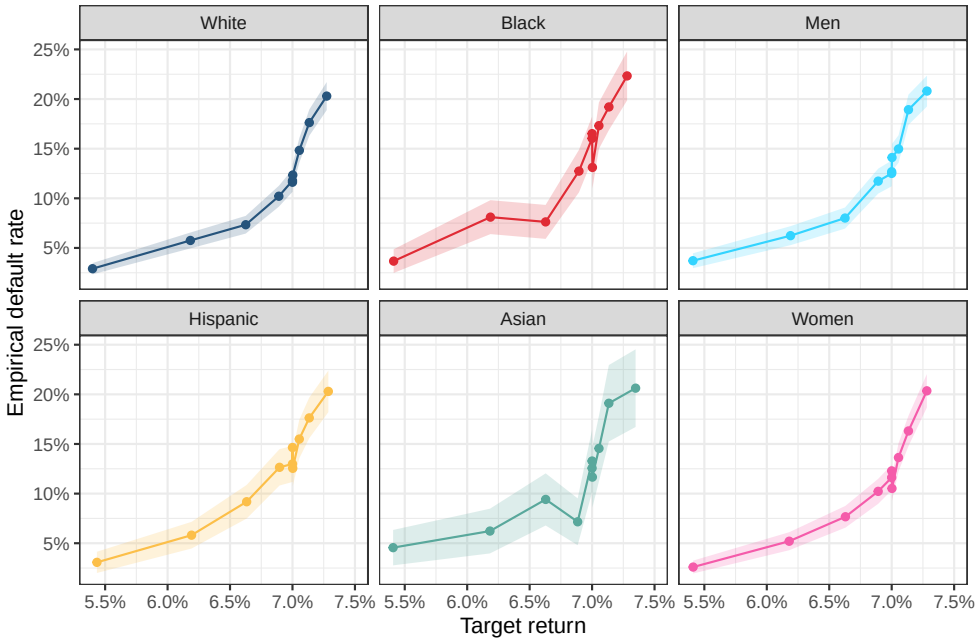

Figure A6: *An empirical assessment of the relationship between target returns and default risk. Curves were generated by computing the average target return and default rate within each group deciles of target return. The shaded area shows the 95% confidence interval of our estimates. Across groups, target default risk exhibits a strong, positive correlation with target returns. These results were computed using the set of funded borrowers.*

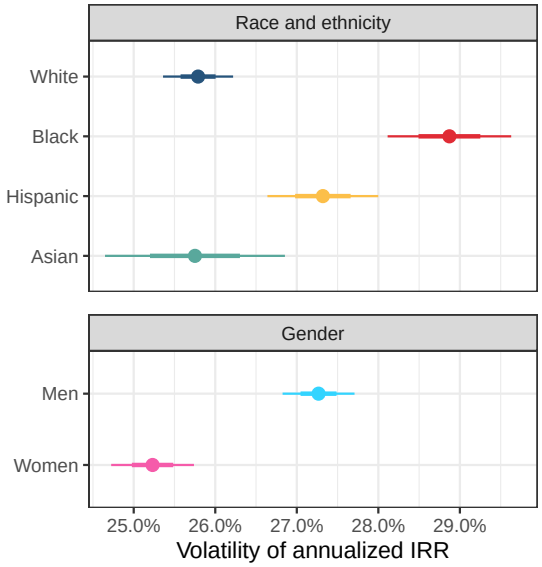

Figure A7: *Volatility of loan IRRs computed across race and gender groups. Black and male borrowers tend to be higher risk groups, as exhibited by their higher volatility of returns. We show the 68% confidence interval (thick bar) and the 95% confidence interval (thin bar) of the volatility estimates. These estimates were computed using the set of funded borrowers.*